the
WINNOWER

CHEMISTRY

Steps to an open tool for pK_a-prediction

SVEN KOCHMANN

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

DATE RECEIVED:
June 10, 2015

,

Last updated on Tuesday, September 16, 2014
<!-- coins metadata inserted by kblog-metadata -->
Update, 16th Sep, 2014: Started an writeLatex-document on this topic.

Is there an open source pKa or LogD tool available?. This question was asked 4 years ago. Still, there is no really good tool available. This post is about my thoughts about creating one.

<!--more-->

So, what do we need for an open tool for pKa-prediction? First, an open database is required to which raw titration data sets can be submitted. The pKa-values of these data sets have to be determined (including statistics) and then linked to the structure and moieties of the corresponding molecules. Finally, methods and models for pKa-prediction based on and/or trained with this data have to be developed.

Acquiring of titration data
In order to get a lot of raw data Chris Swain suggested to let undergraduate students determine pKa-values and put the results in an open database. Although this is a good idea, it requires some preparatory work.

pKa-values are thermodynamic variables, which means that they depend on many environment parameters such as temperature, pressure, ion strength, etc. Also, the concentration (and the grade) of the used substances and solvents (water, DMSO) play a role as well as the used instruments (accuracy of pH-meter, thermometer). Theoretically, the stirring speed and the waiting time between adding a drop and readout of the pH-meter plays a role. What about special cases such as when the solubility or handling strongly depends on the pH (e.g. precipation on low pHs or handling oily substances) or when the pKa is over 12? Hence, standard protocols (plural!) are needed for performing the titration of substances and all recorded data has to be submitted.

Someone could object that using a group of inexperienced students will result in bad data sets. How would you control that the data submitted to the database is ‘good’ (whatever that means) and accurate? I think, statistics will compensate inaccuracy of individuals in this case. If 1000 data sets of students for one compound are submitted – it is nearly impossible that they did all the same random errors. Of course, there could be an systematical error. This should be out of question with a good standard protocol, though. On top of that, a supervisor/instructor who wants to support a project like this by contributing data to it, will have his eye on the students and their performance (I would!).

Update, 6th Sep, 2014: These protocols should be developed by the students themselves. As Anne said in the comments: ‘ Let´s give people the time and experience they need to become data curators, like JCB did with their students. Said that, of course, it should have some quality indicators. Why not discuss this and elaborate with the students?'

Finally, – and this is the reason why I really like this idea from Chris – it will motivate students performing their experiments with great care if they know it has a value. Think back at your lab courses in the 1st, 2nd, or 3rd term/semester/year. You did your experiments, produced some data, the instructor/supervisor reviewed it, and literally … threw it away (or you did it). Reproduction of experiments had no value except practicing and passing the lab course. Wouldn't it be more satisfying for a student if he knew his data was used in 'real' research and that he contributed to it? I think it would!

Setting up a database

Technically, it does not really matter what sort of database one uses for saving the data from above. Only the interfaces for submitting and accessing the data are important. First, people should not have to pass extra hurdles (e.g. logins) when they want to contribute data. It should be easy, straightforward, and rewarding!

Secondly, the data in the database can serve as starting point/base for many projects. It can be used not only for pKa-prediction but for safety reasons (could use it in Beryllium10, for example) or in metabolism/toxicological research projects or as base for semi-empirical quantum chemistry methods or, of course, 'just' as a reference. Thus, having such a database (with many, extensive data sets) is already a great treasure!

Update, 6th Sep, 2014: A wiki as used by the Open Notebook Science Challenge seems like a good option. However, the problem would be to establish the 'no peeking'-rule as suggested by Peter (see comments). The open data principle automatically eliminates this, IMHO. Everyone has access to the data, so everyone can peek. Maybe this should be discussed with the students when introducing them to this?

pKa determination and linking to molecule structure

It is simple to automatically extract pKa-values from a titration plot as long as the type and numbers of acid/base groups are know. Think at the very linear titration plot of citric acid, for example. If you know it is the result of three deprotonation you can easily readout the data and link it to the corresponding moieties.

Actually, this could be done when submitting titration data. The user is presented with the results of the automatical determination, draws the structure of the compound (if not already present), and connects the values to the corresponding protons. On top, the user can choose to manually set, add, or delete pKa-values if the algorithm fails partly or completely – and so train the algorithm. There will be more than one data set for each compound, which gives you the ability to use statistical tools.

pKa prediction

This is the interesting part. It is straightforward, though. Now, that an extensive base set exists (we assume that we have the database from above filled already in), you can develop models based on it and methods, which can be trained by it! Since the data and the database is open and free, just everyone can use it and try to develop something. Of course, although that might sound easy, it is still challenging.

There are all sort of problems with prediction. For instance, symmetry! Think of citric acid again, which has three pKa-values instead of just two. Also, IMHO a good pKa-prediction tool should also be able to predict the dissociation constant of every proton in a molecule such as of aldehydes or just a usual C-H bond. This could be tricky, though.

My idea for a GUI version of a prediction tool would be something very simple. Let the user draw a structure (or load a file or a structure-key) and the tool predicts every pKa-value for every proton present. Nothing more needed (saving the data of course).

It becomes more clear that this is not easy if you look at some implementations.ToxPredict calculates a pKa of 5.21 (instead of about 10) for phenol. Also, it gives you just one pKa for citric acid. Additionally, it does not link this values to any moieties in the molecule.

If you search the literature, you will find all sort of prediction models forpredicting values for proteins, alcohols, and carboxylic acids. And, you will find papers about software such as Epik, which is part of the proprietary Schrödinger suite. Nothing open, though (except AMBIT, which is used by ToxPredict).

Conclusion

The key for developing new and better models and methods, and in turn a prediction tool, is an extensive base data set. Optimally, a free and open data set, everyone can use, which is well documented. So, let us create one!

My two cents.

Update: Remarks (26th Aug, 2014)

- Alex M. Clark suggested to associate names with the data records for professional integrity.
- He also suggested to represent deprotonation as a reaction, rather than just a molecule and a number.
- Marcus Schmid suggested to include some sort of 'ranking' to implement some sort of competition ('University of Somewhere has already submitted 100 data sets? We have to do something to keep up!').

<!-- kcite active, but no citations found -->
<!-- kcite-section 470 -->